Statistical Learning Assignment 1

Sherry Usman

April 22, 2024

1 Part A: Supervised Learning

1. When considering the data generating function, we can see that the relationship between Y and X is non-linear as it includes polynomial terms x_1^2 and x_1^3 and also includes logical conditions like I($x_2 < 0$) and I($x_2 > 3$).
When considering the data-generating function and keeping in mind that we are looking at predictors from X_1 to X_6, K Nearest Neighbors (KNN) might work better as it is a non-parametric supervised learning method that does not make any underlying assumptions about the relationship between the predictors and outcome variable and can capture complex non-linear relationships. Thus it is much more flexible and experiences lower variance (does not overfit) and performs better on unseen data. However, KNN may suffer from a higher bias when the outcome does not capture the relationship between predictors and outcome variable accurately. Lasso Regression on the other hand operates under the strict assumption that the relationship between the predictor and outcome variable is linear and therefore it suffers in this case when the relationship between X and Y i clearly non-linear.
It is important to also note that a reduction of dimensions from X_1 to X_{203} to X_1 to X_6 also has an impact on performance as when excluding irrelevant predictors from X_7 onwards, KNN works better as the lack of noisy variables KNN the distance between datapoints is much more meaningful and accurately captures the relationship between X and Y. Thus the removal of irrelevnat predictors prevents the model from overfitting to noisy data, leading to a lower variance but a much higher bias when testing the model on unseen data.

2. When considering the full set of predictor variables from X_1 to X_{203}, lasso regression might perform better as it tends to be less sensitive to noise and actively reduces the contributions of noisy variables/variables. It does this by adding a regularisation term L_1 to the noise function. The inclusion of this term L_1 allows the lasso regression to be much more robust in dealing with noisy variables and prevents overfitting as it only focuses on relevant predictors. However this may come at the cost of a high bias. On the other hand KNN which relies on the similarity between data-points to find the predicted outcome may suffer with the inclusion of noisy variables as the increased dimensions make it much more computationally expensive to find the nearest neighbor and the distance between neighbors becomes less meaningful. Thus KNN may suffers from overfitting and have higher variance, capturing the noise in data rather than the underlying pattern.

3a. In this question I conduct a 10-fold cross validation for K Nearest Neighbors to find the optimal value of k on the dataset with limited predictor variables. The optimal value of k is the number of neighbors considered for the classification of a specific data point that produces the highest accuracy (most often correctly assigns a datapoint to a particular class/group). A 10-fold cross-validation splits the data randomly into 10 groups/mini-batches and then tests on the first set (considered the validation set) and trains on the remaining sets. This procedure is repeated 10 times for each group and then the mean squared error obtained is then averaged. The 10-fold cross validation results on the training data show that a k=15 produces the highest accuracy. We then implement a KNN with k=15 and find that it has an accuracy of 0.709 (or 70.9%).

3b. In this question I conduct a 10-fold cross validation for Lasso Regression to find the optimal value of lambda λ (the tuning parameter which determines the shrinkage penalty). The results of the cross-validation on the training data show that $\lambda = 0.00463$ produces the highest accuracy. We then

Methods	Number of Predictor Variables	Accuracy (in %)
K-Nearest Neighbors	6	70.9
Lasso Regression	6	66.7
K-Nearest Neighbors	203	56.9
Lasso Regression	203	66.8

Table 1: Comparison of Accuracy between K-Nearest Neighbors and Lasso Regression with Different Numbers of Predictor Variables.

implement a Lasso Regression with λ =0.00463 and find that it has an accuracy of 0.667 (or 66.7%).

3c. From our results we can see that we were right about our initial assumption that KNN would perform better than Lasso Regression. This is because KNN does not make strict assumptions of a linear relationship between X and Y and the removal of noise variables makes the distance between neighbors much more meaningful and helps correctly assign a datapoint to a class. Comparatively, Lasso Regression performs comparatively worse as it tends to assume a linear relationship between predictor and outcome variables and suffers when this relationship is non-linear as shown by the equation in *GenerateSDS.R*.

4a. In this question I conduct a 10-fold cross validation for K Nearest Neighbors to find the optimal value of k on the dataset with the full dataset including all predictor variables from X_1 to X_{203}. The 10-fold cross validation results on the training data show that a k=13 produces the highest accuracy. We then implement a KNN with k=13 and find that it has an accuracy of 0.5686 (or 56.9%).

4b. In this question I conduct a 10-fold cross validation for Lasso Regression to find the optimal value of lambda λ (the tuning parameter which determines the shrinkage penalty). The 10-fold cross validation results on the training data show that $\lambda = 0.0129$ produces the highest accuracy. We then implement a Lasso Regression with λ =0.0129 and find that it has an accuracy of 0.668 (or 66.8%).

4c. From our results in Table 1 it is evident that the KNN performs better than Lasso Regression when the number of predictor variables is limited. This is because KNN is a non-parametric method that does not make assumptions about the relationship between the predictor and outcome variables. Furthermore, the removal of noisy variables ensures the the distance in datapoints is more meaningful and more accurately represents the relationships between X and Y. Lasso Regression on the other hand assumes a linear relationship between the predictor and outcome variables and thus suffers in this case as the equation in GenerateSDS.R shows that y is made up of polynomial terms such as x_1^2 and x_1^3. On the other hand, Lasso Regression performs better when the number of predictor variables is large. This is because Lasso Regression includes a regularisation term L_1 that minimises/penalises the contribution of noisy coefficients and this prevents high variance from overfitting to the data. Thus Lasso Regression is more robust with larger datasets that contain more noise. We can see this in table 1 as the inclusion of predictor variables from X_7 to $X_2$03 has a limited effect on the accuracy of the Lasso Regression model (0.1%). On the other hand KNN cannot actively reduce noisy contributions like Lasso Regression and thus suffers as the inclusion of noise makes the distance between datapoints much less meaningful. The model then begins to overfit to the noise rather than capturing the underlying relationships between Xs and Y. This is seen by a stark drop in accuracy from 70.9% to 56.9%.

2 Part B: Unsupervised Learning

For the second section we explore the *data.US.csv* dataset. This dataset that consists of 1000 individuals who have been measured on 30 personality variables ranging from V2 to V31. The aim of this section is to find groups with similar personality types.

1. Dimensionality reduction may be beneficial in this particular dataset as this dataset has a large number of variables. Shrinking the number of variables into dimensions/components may make the contributions of each component much more meaningful (as it represents a larger fraction of the

dataset) and prevent overfitting to many less meaningful variables. Furthermore, the individuals are separated into different personality types based on different personality variables. However, when reading the list of emotions we can see that there is some similarity between some personality variables and such variables could can be grouped with others to get one component that is more meaningful such as **anxiety** and **self-consciousness** and **order** and **self-discipline** contributes more to the variance of the dataset. Furthermore, we can assume that the contributions from variables such as **feelings** and **aesthetics** might not be as interesting as contributions from variables such as **anxiety** and **gregariousness**. Dimensionality reduction solves this problem by making the list of dimensions much more concise and efficient.

2. **PCA:** To find the ideal number of Principal Components we first conducted a Principal Component Analysis (PCA) using the `prcomp()` function in R and plotted the eigenvalues of each component in a bar graph. As seen from the figure the first 5 components have eigenvalues greater than 1.

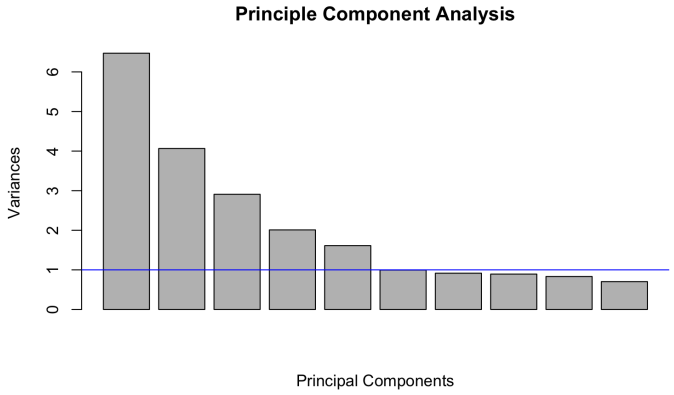

Figure 1: This figure represents the results from the Principal Component Analysis of the *data.US.csv* dataset. The blue line represents a cut-off point for all eigenvalues less than 1.

Since Kaiser rule states that only the principal components whose eigenvalues exceed 1 should be retained, the first 5 components are selected for PCA. the eigenvalue has to be more than 1, the first 5 components should be selected from PCA.

Screeplot: To further validate my hypothesis I created two screeplots to help visualise the proportion of variance each principal component contributes to the dataset. This is shown in Figure 2. Using the elbow rule, we found that the ideal number of principal components to choose was 5.

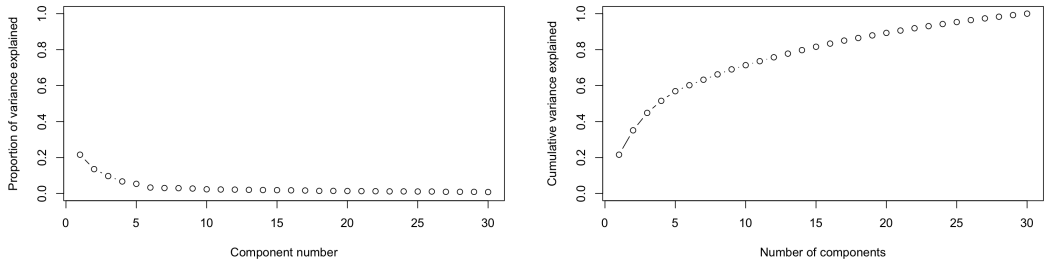

Figure 2: The plots represents the proportion of variance and cumulative variance explained by each principal component in the *data.US.csv* dataset.

Velicer's MAP test: Using the `MAP()` function in R, we conducted the Velicer's MAP test and found that the ideal number of components to retain is 5.

Horn's Parallel Analysis: Using the `paran()` function in the paran library in R, we conducted Horn's Parallel Analysis and found the ideal number of components retained is 5.

3. To understand how each much each variable contributes to each principal component we used the `fviz.contrib()` function from the facto-extra library and producted the graphs shown in Figures 3, 4 and 5, showing the percentage of contributions of each variable to each principal component.

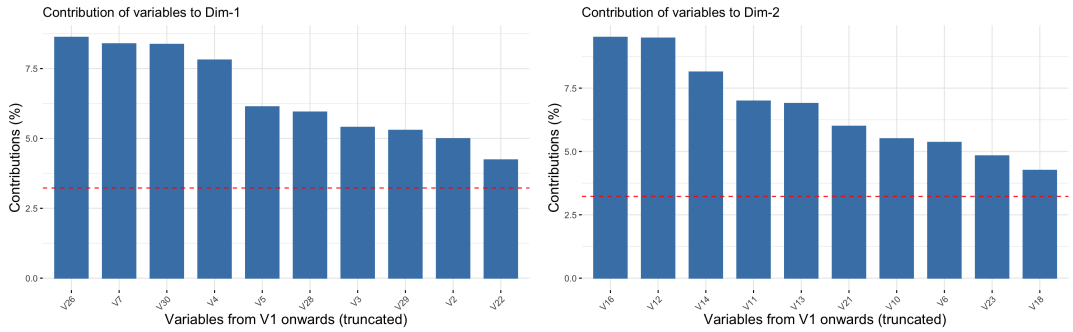

Figure 3: This figure represents the proportion of variance explained by each principal component in the *data.US.csv* dataset.

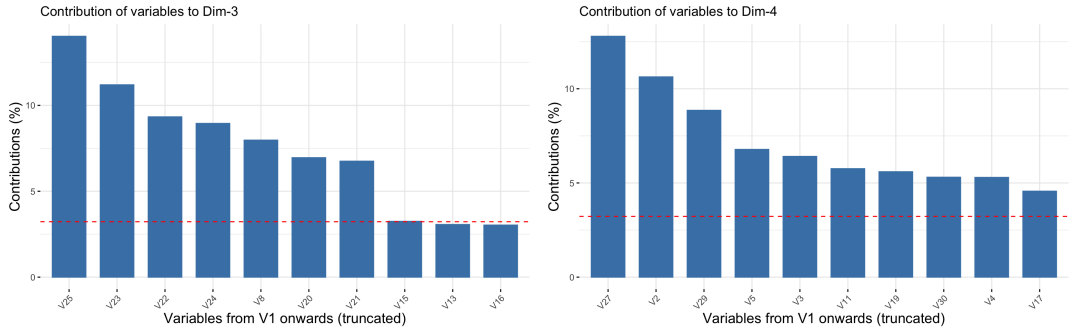

Figure 4: This figure represents the proportion of variance explained by each principal component in the *data.US.csv* dataset.

Thus we can see that PC1 corresponds to V26 (competence), V7 (vulnerability), V30 (self-discipline), V4 (depression), V5 (self consciousness), V28 (dutifulness), V3 (angry hostility), V29 (achievement-striving), V2 (anxiety) and V22 (altruism).
PC2 corresponds to V16 (feelings), V12 (excitement-seeking), V14 (fantasy), V11 (activity), V13 (positive emotions), V10 (assertiveness), V6 (impulsiveness), V23 (compliance) and V18 (actions).
PC3 corresponds to V25 (tender-mindedness), V23 (compliance), V22 (altruism), V24 (modesty), V8 (warmth), V20 (trust) and V21 (straightforwardness).
PC4 corresponds to V27 (order), V2 (anxiety), V29, V5 (self consciousness), V3 (angry hostility), V11 (activity), V19 (values), V30 (self-discipline), V4 (depression) and V17 (ideas).
Lastly, PC5 corresponds to V15 (aesthetics), V18 (actions), V9 (gregariousness), V12 (excitement-seeking), V31 (deliberation), V13 (positive emotions) and V14 (fantasy).

4. Different methods can be used to find the cumulative variance explained from the principal components. Velicer's MAP test shows that the first 5 principal components explain 57% of the total variance, while the summary of our PCA results show that the cumulative variance captured by the principal components is 56.8%. The difference in both methods is negligible.

5. Different statistical techniques can be used to group participants, for example K-Means and Hierarchical Clustering. Both k-Means and Hierarchical clustering are unsupervised learning techniques

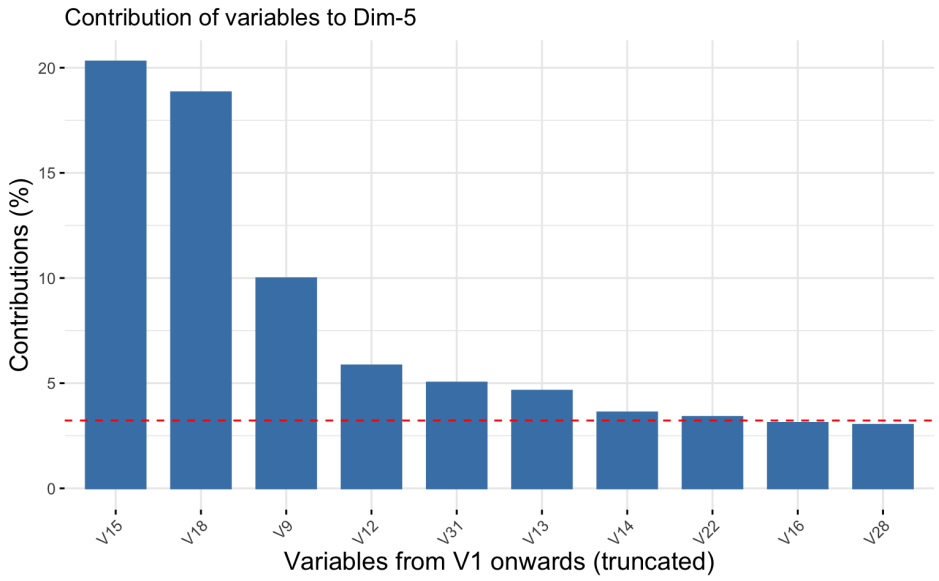

Figure 5: This figure represents the proportion of variance explained by each principal component in the *data.US.csv* dataset.

that can group objects/points that share similar features, in this case individuals that share similar personality traits. K-Means offers a number of advantages that makes it a suitable method for this question. For example, it guarantees convergence towards a stable number of clusters. Furthermore, it is easy to implement in R and can be specialised to different cluster shapes and sizes like ellipse. Hierarchical clustering on the other hand is also an appropriate statistical technique and it offers a number of advantages as it does not require a pre-defined number of clusters and easily summarises data into a dendrogram.

6. Before implementing K-Means clustering it is important to calculate the optimal number of clusters/groups for our principal components. Then we can implement K-Means clustering with this optimal value of k. The figure below shows that the optimal value of k is 5. That means the principal components can be optimally divided into 5 groups with the least amount of overlap and mis-assigning datapoints. The results of K-Means Clustering are shown in Figure 7.

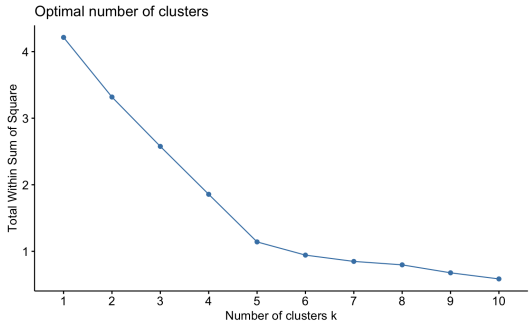

Figure 6: This figure represents the optimal number of clusters for K-Means Clustering of principal components and their clusters

It is also better to implement another unsupervised learning method to ensure that our results are valid. For this I chose hierarchical clustering using the agglomerative/bottom-up method. In this method, each datapoint is initially assigned as a cluster and cluster closer to each other are repeatedly joined until only one cluster remains. The results of hierarchical clustering are shown in Figure 8.

7. Each group contains 5 to 8 variables in both cases.

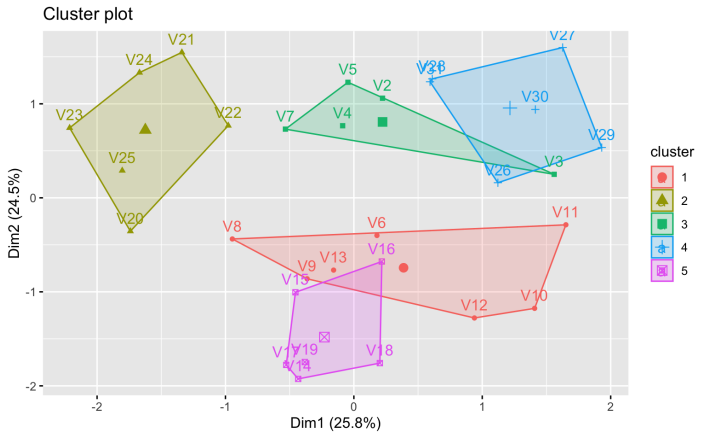

Figure 7: This figure represents the result of K-Means Clustering of principal components

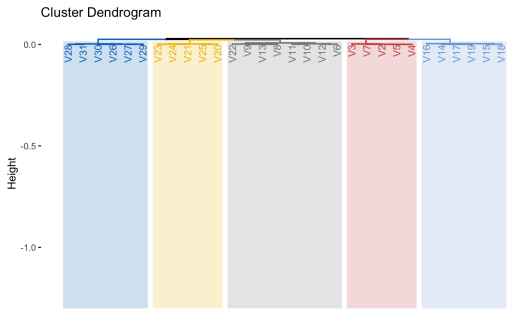

Figure 8: This figure represents the result of Hierarchical clustering of principal components

8. The two unsupervised learning methods have their differences. In K-Means clustering the number of clusters is pre-defined by a value k while hierachical clustering, indicative of its name, builds a hierarchy of clusters (top-down/ divisive) or (bottom-up/agglomerative) without having a fixed number of clusters beforehand. In this case one can use the dendrogram to stop at any number of clusters they deem suitable. Thus it adds a degree of flexibility. K-Means does not make assumptions about the underlying distribution of data, making it suitable for all kinds of data. Furthermore, because K-Means clustering uses a nearest neighbors approach it is often less computationally expensive than hierarchical clustering which often takes an agglomerate or divisive approach. The agglomerative approach for hierarchical clustering has a time complexity of $O(n^3)$ and requires $\Omega(n^2)$ memory where n is the size of the dataset, which makes it extremely slow for larger datasets. However, it works well here as we implement it on only the rotational loadings of 5 principal components.

9. K-Means clustering gives us five groups / personality types. Group 1 is defined by the words: impulsiveness, warmth, gregariousness, assertiveness, activity, excitement-seeking and positive emotions. This aligns with the a personality trait in the Big Five Personality Traits called extarversion [(Theresa A. Morgan, 2013)]. Extraverted people are often more sociable, active and outgoing, constantly seeking excitement. They are also more assertive in their decision making and value action.
Group 2 is defined by the words: trust, straightforwardness, altruism, compliance, modesty and tendermindedness. This aligns with the a personality trait in the Big Five Personality Traits called agreeableness [(Theresa A. Morgan, 2013)]. An agreeable person is is kind, compassionate, agreeable and willing to help others while also remaining modest.
Group 3 is defined by the words anxiety, angry hostility, depression, self-consciousness and vulnerability. This aligns with a personality trait in the Big Five Personality Traits called neuroticism [(Theresa A. Morgan, 2013)]. Such a person is high-strung and suffers from anxiety and are constantly hyperaware of their actions and appearance in social situations.
Group 4 is defined by the words competence, order, dutifulness, achievement-striving, self-discipline and deliberation. This aligns with a personality trait in the Big Five Personality Traits called conscientiousness [(Theresa A. Morgan, 2013)]. A conscientious person is an individual who values order and discipline and is constantly exhibits that in their daily lives, through working to achieve their goals, showing up for the people around them and staying committed to their relationships and goals.
Lastly group 5 is defined by the words fantasy, aesthetic, feelings, ideas, actions and values. This aligns with a personality trait in the Big Five Personality Traits called openess [(Theresa A. Morgan, 2013)]. Such a person is open-minded, curious, adventurous and emotionally expressive. They tend to have a strong appreciation of the arts in its many forms like poetry, painting and creative-writing.

References

Michael S. Chmielewski Theresa A. Morgan. Five-factor model of personality. 2013. URL https://link.springer.com/referenceworkentry/10.1007/978-1-4419-1005-9_1226#:~:text=It%20consists%20of%20five%20main,)%2C%20Agreeableness%2C%20and%20Conscientiousness.

3 Appendix

Principal Component	Variables	Meaning
Cluster 1	V6	Impulsiveness
	V8	Warmth
	V9	Gregariousness
	V10	Assertiveness
	V11	Activity
	V12	Excitement Seeking
	V13	Positive emotions
Cluster 2	V20	Trust
	V21	Straightforwardness
	V22	Altruism
	V23	Compliance
	V24	Modesty
	V25	Tender-mindedness
Cluster 3	V2	Anxiety
	V3	Angry hostility
	V4	Depression
	V5	Self-consciousness
	V7	Vulnerability
Cluster 4	V26	Competence
	V27	Order
	V28	Dutifulness
	V29	Achievement-striving
	V30	Self-discipline
	V31	Deliberation
Cluster 5	V14	Fantasy
	V15	Aesthetics
	V16	Feelings
	V17	Ideas
	V18	Actions
	V19	Values

Table 2: A table of the results of K-Means Clustering, showing the different clusters and their personality traits